Project 1: Regression Model to Predict the Value of Homes in the Sindian District, New Taipei City, Taiwan.

Due before midnight on September 21, 2020

For our first machine learning project you will develop a model to predict the value of homes in an area of Taiwan based on a real set of data from 2018[1]. The data contains 6 features (x's) along with the value (y) of the house per unit area (in multiples of 10,000 New Taiwan Dollars/Ping, 1 Ping = 3.3 meters squared).

x_1=the transaction date (for example, 2013.250=2013 March, 2013.500=2013 June, etc.)
x_2=the house age (unit: years)
x_3=the distance to the nearest MRT Rapid Transit subway station (unit: meters)
x_4=the number of convenience stores nearby (integer)
x_5=the geographic coordinate, latitude. (unit: degrees)
x_6=the geographic coordinate, longitude. (unit: degrees)

Your project is to develop a regression model that will predict the value of a home per ping based on these six features. (A 25 means that the value is 250,000 New Taiwan Dollars/ping or about US$86/SqFoot). Your data set is named REData.txt and is formatted as follows:

first line: integer number of lines of training data (m = 414), tab, integer number of features (n = 6)

all other lines: values of input features, tab separated, followed by the value (y) of the house per unit area (in order listed above).

Example:

m n
$x^{(1)}_1$ $x^{(1)}_2$ $x^{(1)}_3$ . . . $x^{(1)}_n$ $y^{(1)}$
. . .
$x^{(m)}_1$ $x^{(m)}_2$ $x^{(m)}_3$ . . . $x^{(m)}_n$ $y^{(m)}$

In order for us to evaluate your model it must follow strict input and output requirements. Your Python program must run in Spyder.

- It should first prompt the user for a Training Set file name that is formatted as:
  - First line: integer number of lines of training data (m), tab, integer number of features (n)
  - all other lines: values of input features, tab separated, followed by the value of the house (in order listed above).
- Next: Prints out values for all weights and final J for the training set. All output should be clearly labelled.
- Next: Prompts the user for the name of either a validation data set or a test data set. Same file format as the training data.
- Next: Prints out J for and adjusted R^2. All clearly labelled.

[1] Yeh, I. C., & Hsu, T. K. (2018). Building real estate valuation models with comparative approach through case-based reasoning. *Applied Soft Computing*, 65, 260-271.

Model Development

First, you should randomize the original data set. (This happens only once.)

After randomization you can divide your data into a training file (248 data sets), a validation file (83 data sets) and a test file (83 data sets). You should test three different models.

The data as given where each line of data is $x_1, x_2, x_3, x_4, x_5, x_6, y$

Squaring the data so that each line of data is $x_1^2, x_2^2, x_3^2, x_4^2, x_5^2, x_6^2, y$

Pairing the original data with the squared data: $x_1, x_1^2, x_2, x_2^2, x_3, x_3^2, x_4, x_4^2, x_5, x_5^2, x_6, x_6^2, y$

The same regression program should be capable of running any set of data without modification as long as the file is formatted correctly. The important thing to remember is that the different models should be represented entirely by the data file! Your program should be constructed so that it will work for any training set that lists, m, n on the first line and data sets on each following line.

What to turn in (uploaded to Canvas as a single zipped file named yourlastname_yourfirstname)

1. A pdf file containing:
    - A table describing your different models and giving your J values for test and validation sets for all models, and J and R^2 values for the test set **for the best model based on your validation results**.
    - A list of your final weights.

Example of table:

Model	J_{train}	$J_{validation}$	J_{test}	Adjusted R^2_{test}
Original Data				
Squared Data				
Original + Squared Data				

2. Your Python program named: yourlastname_yourfirstname_P1.py
3. Your Final Training set named: yourlastname_yourfirstname_Train.txt
4. Your Final Validation set named: yourlastname_yourfirstname_Valid.txt
5. Your Final Test set named: yourlastname_yourfirstname_Test.txt

DO NOT ASSUME that any files will exist other than the ones you submit when we test your program. In particular, the original data file will not be available.

Note: I suggest implementing your regression program using the Normal Equation instead of Gradient Descent, but either are acceptable.